This is the Pre-Published Version.

This version of the proceeding paper has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use(https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms), but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-72120-5_8.

Anatomical Structure-Guided Medical Vision-Language Pre-training

Qingqiu Li[1], Xiaohan Yan[2], Jilan Xu[1], Runtian Yuan[1], Yuejie Zhang[1(✉)], Rui Feng[1(✉)], Quanli Shen[3], Xiaobo Zhang[3], and Shujun Wang[4,5]

[1] Fudan University [2] Tongji University
[3] Children's Hospital of Fudan University
[4] The Hong Kong Polytechnic University
[5] Research Institute for Smart Ageing

Abstract. Learning medical visual representations through vision-language pre-training has reached remarkable progress. Despite the promising performance, it still faces challenges, i.e., local alignment lacks interpretability and clinical relevance, and the insufficient internal and external representation learning of image-report pairs. To address these issues, we propose an **A**natomical **S**tructure-**G**uided (ASG) framework. Specifically, we parse raw reports into triplets <*anatomical region*, *finding*, *existence*>, and fully utilize each element as supervision to enhance representation learning. For anatomical region, we design an automatic anatomical region-sentence alignment paradigm in collaboration with radiologists, considering them as the minimum semantic units to explore fine-grained local alignment. For finding and existence, we regard them as image tags, applying an image-tag recognition decoder to associate image features with their respective tags within each sample and constructing soft labels for contrastive learning to improve the semantic association of different image-report pairs. We evaluate the proposed ASG framework on two downstream tasks, including five public benchmarks. Experimental results demonstrate that our method outperforms the state-of-the-art methods.

Keywords: Representation Learning · Medical Vision-Language Pre-training · Contrastive Learning · Anatomical Structure

1 Introduction

In recent years, vision-language pre-training (VLP) has achieved remarkable success [8, 10, 11, 15]. These models are trained using millions of web image-text pairs by matching images to their corresponding captions without the need for manual labels. In the medical domain, this paradigm also gains increasing attention. The models trained on paired image-reports benefit a broad spectrum of downstream medical image understanding tasks. Among them, ConVIRT [25] first used contrastive learning as a proxy task for biomedical data processing. Building upon this, GLoRIA [5] and MGCA [19] further employed local-level alignment to bolster model performance. Moreover, MedKLIP [22] and KAD

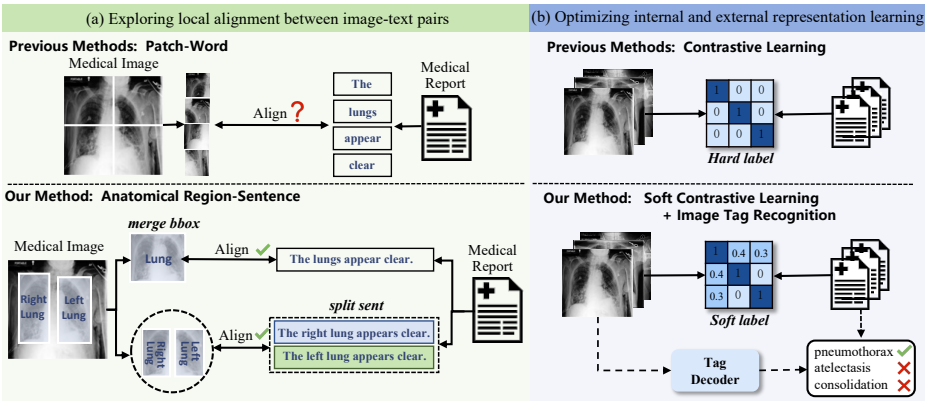


Fig. 1. Two limitations of existing methods: (a) lack of interpretability and clinical relevance and (b) insufficient representation learning of image-report pairs; and our corresponding improvement.

[24] incorporated additional domain knowledge to guide better representation learning. MRM [27] encouraged the model to pay attention to low-level features through mask reconstruction, enhancing model's utility in downstream dense prediction tasks.

Despite the advancements made in the medical VLP scenario, we identify two inherent limitations of existing methods that remain unresolved. **1) Lack of interpretability and clinical relevance.** Previous methods [5,19] explored local alignment between image-text pairs at patch-word level or patch-sentence level, which lacks semantic and clinical correspondence. For example, local patches extracted from X-rays represent neither lesion area nor anatomical organs. In addition, decomposing the whole sentence into individual words results in a substantial loss of semantic context. Consequently, seeking alignment between clinically irrelevant visual patches and semantically unrelated words diminishes interpretability and disturbs the model optimization. **2) Insufficient representation learning of image-report pairs.** Compared to natural image captions, medical reports are generally longer and contain richer medical knowledge, posing additional challenges for comprehensive report understanding. However, most previous approaches [5,19] still rely on simple encoding of raw reports to extract text features, which fails to fully understand the entire report and capture the key sections describing lesions. KAD [24] and MedKLIP [22] addressed this issue by using entity recognition tool to obtain structured reports, thereby improving the supervision of text for significant image representations. Nevertheless, they neglected the semantic analysis between different pairs and employed hard labels in contrastive learning, leading to a considerable number of false negatives.

In this paper, we propose a novel **A**natomical **S**tructure-**G**uided (ASG) framework to introduce anatomical knowledge into medical VLP, thus achieving clinically reliable representation learning. We parse raw reports into triplets

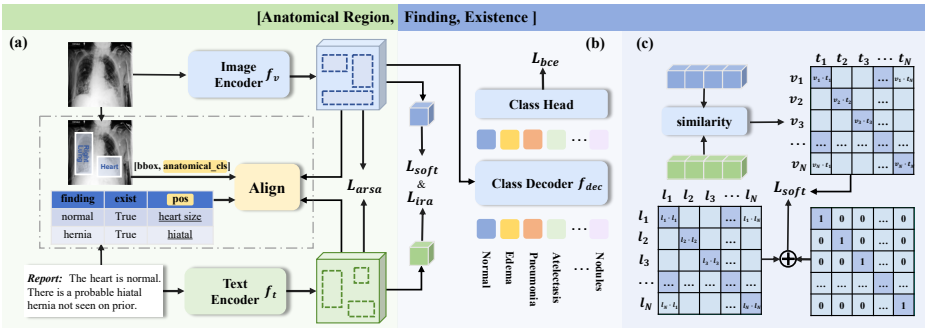


Fig. 2. Overview of our ASG framework. (a) Exploring local alignment between image-text pairs through anatomical region-sentence alignment. (b) Optimizing internal representation learning by applying an image-tag recognition decoder to associate image features with their respective tags. (c) Optimizing external representation learning by constructing soft labels for contrastive learning to mitigate false negatives.

<*anatomical region*, *finding*, *existence*>, and utilize each element as supervisory information. Specifically, under the guidance of radiologists, we design an automatic anatomical region-sentence alignment paradigm, which aligns with radiologists' reading workflow and enhances interpretability. Furthermore, we simultaneously focus on the internal and external semantic features of image-report pairs, utilizing an image-tag recognition decoder to associate image features with their respective tags and constructing soft labels for contrastive learning to mitigate false negatives. Extensive experiments have been conducted on two downstream tasks, including five public benchmarks, demonstrating that our method outperforms other state-of-the-art methods.

2 Methodology

As illustrated in Fig. 2, we propose a novel **A**natomical **S**tructure-**G**uided (ASG) framework for medical VLP, which consists of three parts: Image-Report Alignment (IRA), Anatomical Region-Sentence Alignment (ARSA), and Internal and External Representation Learning (IERL). In this section, we first introduce the visual and text encoding, followed by elaborating on each part in detail.

Given a batch of N image-report pairs $(\mathbf{I}, \mathbf{R}) = \{(I_i, R_i)\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ denotes the image and $R_i = \{s_i^j\}_{j=1}^{M_S}$ is the report containing M_S sentences. For each image, we utilize an image encoder f_v to get a sequence of M_Z feature representations $\mathbf{z}_i^v = f_v(I_i) \in \mathbb{R}^{M_Z\times d}$. The global visual representation is computed by averaging the dense features, *i.e.*, $\mathbf{v}_i = \text{Avg}(\mathbf{z}_i^v) \in \mathbb{R}^d$. For each report, we employ a text encoder f_t to encode it into a sequence of sentence tokens $\mathbf{z}_i^t = f_t(R_i) \in \mathbb{R}^{M_S\times d}$ and a global report feature $\mathbf{t}_i \in \mathbb{R}^d$.

2.1 Image-Report Alignment (IRA)

To achieve the image-report alignment, we enforce the paired global image and report representation to be close in the feature space by employing the instance-level contrastive learning. Two non-linear projection layers are applied to obtain normalized lower-dimensional embeddings, *i.e.*, $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{t}}_i$. After that, we calculate the image-to-report similarity $\mathbf{p}(I_i, R_j) = \frac{\exp(\langle \hat{\mathbf{v}}_i, \hat{\mathbf{t}}_j \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \hat{\mathbf{v}}_i^t, \hat{\mathbf{t}}_k \rangle / \tau)}$, and report-to-image similarity $\mathbf{p}(R_i, I_j) = \frac{\exp(\langle \hat{\mathbf{t}}_i, \hat{\mathbf{v}}_j \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \hat{\mathbf{t}}_i, \hat{\mathbf{v}}_k^v \rangle / \tau)}$, where $\langle , \rangle$ represents the cosine similarity and τ is the temperature hyperparameter. IRA is optimized by the InfoNCE loss [14] to maximize the similarity between paired instances:

$$\mathcal{L}_{ira}^{(v \to t)} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(\mathbf{y}_i, \mathbf{p}(I_i, \mathbf{R})), \ \mathcal{L}_{ira}^{(t \to v)} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(\mathbf{y}_i, \mathbf{p}(R_i, \mathbf{I})), \tag{1}$$

where $\mathrm{H}(,)$ denotes the cross-entropy, $\mathbf{y}_i = \{(y_{ij})\}_{j=1}^{N} \in \mathbb{R}^N$ is the one-hot label with y_{ii} equal to 1 and all other elements equal to 0. The overall objective of IRA can be denoted as $\mathcal{L}_{ira} = \frac{1}{2}(\mathcal{L}_{ira}^{(v \to t)} + \mathcal{L}_{ira}^{(t \to v)})$.

2.2 Anatomical Region-Sentence Alignment (ARSA)

To explore fine-grained local alignment between image-text pairs, we propose an Anatomical Region-Sentence Alignment (ARSA) module by discovering clinical relevance. Specifically, we first extract the anatomical regions from the image and anatomical sentences from the report, and then we provide an automated alignment solution. Based on the aligned anatomical region-sentence pairs, we facilitate contrastive learning to realize Anatomical Region-Sentence Alignment.

Extraction of anatomical sentences. For each report R_i, we employ RadGraph [7] to decompose it into a total of M_T triplets $\{\mathcal{T}_j\}_{j=1}^{M_T}$, where each triplet $\mathcal{T}_j$ is denoted as $<$*anatomical region*, *finding*, *existence*$>$, *e.g.*, $<$*lung*, *pneumothorax*, *exist*$>$. Here, the anatomical region in each triplet belongs to an anatomical set C_{ana} [22, 24]. Notably, each triplet $\mathcal{T}_j$ corresponds to one anatomical-related sentence in the report.

Extraction of anatomical regions. For each image I_i, we use an off-the-shell Faster R-CNN [16] pre-trained on Chest ImaGenome Dataset [23] to get a total of M_A anatomical bbox $\mathbf{A}_i = \{(\mathbf{b}_i^j, \mathbf{a}_i^j)\}_{j=1}^{M_A}$. Here, each bounding box $\mathbf{b}_i^j \in \mathbb{R}^4$ represents an region with anatomical class $\mathbf{a}_i^j \in C_{pre}$ in the image (*e.g.*, *right hilar structures*), and C_{pre} is the set of pre-defined categories [23].

Automatic alignment paradigm. To align each anatomical bbox $(\mathbf{b}_i^j, \mathbf{a}_i^j)$ with one triplet $\mathcal{T}$, the main challenge of building bbox-triplet alignment is two-fold. (1) The mismatch between the size of anatomical set $|C_{ana}| = 50$ and the pre-defined categories of the detector $|C_{pre}| = 29$; (2) The semantic overlap of C_{ana} and C_{pre}, *e.g.*, *lung* defined in C_{pre} is correspond to both *left lung* and *right lung* in C_{ana}. To address the issues, we develop an automated paradigm for strict alignment based on the prior knowledge from experienced radiologists.

In particular, given an anatomical region $pos \in C_{ans}$ and a pre-defined class $\mathbf{a} \in C_{pre}$, we mainly consider three scenarios. (see Supp. Fig. 1)
Scenario 1: *pos* and **a** are identical words or phrases, or they share the same region despite different expressions, *e.g.*, *pos* refers to *right hilar* and **a** is *right hilar structures*. Hence, an exact match between anatomical region *pos* and class **a** can be set.
Scenario 2: If **Scenario 1** is not satisfied, we pair *pos* with an **a** that can encompass it, *e.g.*, *pos* = *right ventricle* and **a** = *cardiac silhouette*. Both **Scenario 1** and **Scenario 2** can be symbolized as:

$$\forall j_P \in \{1, 2, \ldots, M_T\},\ \exists j_A \in \{1, 2, \ldots, M_A\},\ \text{ s.t. } pos_i^{j_P} \rightarrow \mathbf{a}_i^{j_A}. \tag{2}$$

Scenario 3: There exists a one-to-many relationship between *pos* and **a**. For example, when *pos* is *diaphragm unspec* and **a** does not have bbox which can fully encompass the entire diaphragm but only includes bboxes for the *left diaphragm* and *right diaphragm*. Here, we propose two solutions: merging anatomical bboxes or splitting the sentences.

$$\forall j_P \in \{1, 2, \ldots, M_T\},\ \exists j_A, j'_A \in \{1, 2, \ldots, M_A\},\ \text{ s.t. } pos_i^{j_P} \rightarrow \{\mathbf{a}_i^{j_A}, \mathbf{a}_i^{j'_A}\} \tag{3}$$

- Splitting the sentences - let $s_i^{j_S}$ be the matched sentence in the report for the anatomical region $pos_i^{j_P}$, we split $s_i^{j_S}$ into $\widetilde{s_i^{j_S}}$ and $\hat{s_i^{j_S}}$. Thus, the matched pairs can be formed as $(\widetilde{s_i^{j_S}}, \mathbf{a}_i^{j'_A})$, $(\hat{s_i^{j_S}}, \mathbf{a}_i^{j_A})$. *E.g.*, *diaphragm unspec* is split into *left diaphragm* and *right diaphragm*, ensuring a strict correspondence with two bounding boxes.

- Merging anatomical bboxes - the matched pair is constructed by merging two anatomical bboxes: $(s_i^{j_S}, \mathbf{a}_i^{j_A} \cup \mathbf{a}_i^{j'_A})$. *E.g.*, the bboxes for *left diaphragm* and *right diaphragm* are merged to obtain the entire diaphragm region.

After obtaining the local anatomical region-sentence pairs of each sample, we calculate the contrastive learning loss $\mathcal{L}_{arsa}$, which is optimised by InfoNCE loss implemented on the region-sentence level.

2.3 Internal and External Representation Learning (IERL)

In Section 2.2, we decompose the raw report into triplets and focus on anatomical structures to do the local alignment. Similarly, the finding and existence of triplets are crucial for fine-grained matching. We consider them as tags $\mathbf{l}_i = \{l_i^j\}_{j=1}^{M_Q}$ for image-report pairs, and utilize these tags to optimize internal and external representation learning. If the current pair has a disease of class j, then $l_i^j = 1$, otherwise, it is 0. Here, M_Q is the number of disease classes.
Internal representation learning. Internal representation learning aims to discover the relationship between image and tags within each sample. For each image-report pair, we apply an image-tag recognition decoder f_{dec} to associate image features with their respective tags. Specifically, we use the sequence of

encoded visual tokens $\mathbf{Z}_i = \{\mathbf{z}_i^j\}_{j=1}^{M_Z}$ as both key and value, and utilize a collection of disease classes $\mathbf{Q} = \{q^j\}_{j=1}^{M_Q}$ as queries. The classification loss is formulated as follows:

$$\mathcal{L}_{bce} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_Q}\sum_{j=1}^{M_Q}(l_i^j \log f_{dec}(\mathbf{Z}_i, Q) + (1 - l_i^j)\log f_{dec}(\mathbf{Z}_i, Q)). \tag{4}$$

External representation learning. External representation learning aims at improving visual-text alignment by exploring tags to connect different image-report pairs. Previous methods [24] simply adopted hard labels by treating paired texts (reports from the same patient's study) as positives and unpaired texts (reports from other patients' studies) as negatives. Nevertheless, hard labels introduce many false negatives, as reports from different patients could have identical symptoms. Therefore, we explore soft label $\mathbf{p}^{soft}(I_i, \mathbf{R}) = \mathbf{p}^{soft}(R_i, \mathbf{I}) = \{p_{ij}^{soft}\}_{j=1}^{N}$ to capture the deep semantic associations between different pairs, which are constructed based on the cosine similarity between different tags:

$$p_{ij}^{soft} = \frac{\exp(\langle \mathbf{l}_i, \mathbf{l}_j \rangle / \tau)}{\sum_{k=1}^{N}\exp(\langle \mathbf{l}_i, \mathbf{l}_k \rangle / \tau)}. \tag{5}$$

In practice, we use the weighted average of the hard labels and the soft labels as the final label to ensure the training stability and better generalization, which is formulated as $\hat{\mathbf{p}}(I_i, \mathbf{R}) = \hat{\mathbf{p}}(R_i, \mathbf{I}) = (1-\alpha)\mathbf{y}_i + \alpha\mathbf{p}_i^{soft}$. Finally, the Kullback-Leibler (KL) divergence is used as the loss to minimize the distance between the final label and the similarity score $\mathcal{L}_{soft} = \frac{1}{2}(\mathcal{L}_{soft}^{(v\to t)} + \mathcal{L}_{soft}^{(t\to v)})$, where :

$$\mathcal{L}_{soft}^{(v\to t)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}(\hat{\mathbf{p}}(I_i, \mathbf{R})||\mathbf{p}(I_i, \mathbf{R})),\ \mathcal{L}_{soft}^{(t\to v)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}(\hat{\mathbf{p}}(R_i, \mathbf{I})||\mathbf{p}(R_i, \mathbf{I})). \tag{6}$$

Overall objective. We train our ASG framework by jointly optimizing the following losses:

$$\mathcal{L} = \mathcal{L}_{ira} + \mathcal{L}_{arsa} + \mathcal{L}_{bce} + \mathcal{L}_{soft}. \tag{7}$$

3 Experiments

3.1 Experimental Setting

Pre-training Setting We pre-train our framework on MIMIC-CXR [9] and follow previous works to preprocess the dataset. The frontal view of images and the reports with more than 3 tokens are selected to generate 217k image-report pairs. We use ResNet50 [4] and ViT-B/16 [3] as the image encoders, and BioClinicalBERT [1] is the text encoder. Our ASG is trained 50 epochs on 4 RTX 3090 GPUs with a batch size of 72 per GPU. We use AdamW [13] as our optimizer, setting the learning rate to $4e-5$ and the weight decay to $5e-2$. And a linear warm-up and cosine annealing scheduler [12] are applied in the process.

Table 1. Comparison with other SOTA methods on the classification task.

Method	NIH X-ray (AUC)			CheXpert (AUC)			RSNA (AUC)			COVIDx (ACC)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init	52.1	54.6	55.3	56.1	62.6	65.7	58.9	69.4	74.1	50.5	60.3	70.0
ImageNet Init	67.0	67.5	71.6	74.4	79.7	81.4	74.9	74.5	76.3	64.8	78.8	86.3
CNN-based												
ConVIRT [25]	64.9	77.1	80.8	85.9	86.8	87.3	77.4	80.1	81.3	72.5	82.5	92.0
GLoRIA [5]	59.7	74.3	80.0	87.1	88.7	88.0	87.0	89.4	**90.2**	66.5	80.5	88.0
MedKLIP [22]	60.9	74.8	80.1	82.3	85.4	87.3	83.3	86.6	88.1	74.5	83.5	91.3
KAD [24]	78.7	80.7	82.5	87.2	88.6	88.7	86.7	88.7	89.9	73.5	83.0	90.5
MGCA [19]	77.7	80.8	82.6	87.6	88.0	88.2	87.6	88.6	89.8	72.0	83.5	90.5
Ours	77.0	81.0	82.9	87.7	88.2	88.7	87.2	88.8	89.7	77.3	84.8	**93.3**
ViT-based												
MRM [27]	78.0	82.1	83.2	88.5	88.5	88.7	87.2	88.7	89.7	79.0	85.5	92.5
MGCA [19]	78.9	82.1	83.5	**88.8**	**89.1**	**89.7**	**88.6**	**89.5**	90.0	74.8	84.8	92.3
Ours	**79.5**	**82.2**	**83.6**	87.9	89.0	89.0	88.4	**89.5**	**90.2**	**81.3**	**87.0**	**93.3**

Downstream Tasks (1) Medical Image Classification We conduct medical image classification on four representative datasets, NIH ChestX-ray [21], CheXpert [6], RSNA [18], and COVIDx [20]. We use the linear probe classification setting to evaluate the transfer ability of our pre-trained image encoder. **(2) Medical Semantic Segmentation** We evaluate the performance of our framework for medical semantic segmentation on SIIM [2] and RNSA [18] datasets. We use the pre-trained ResNet50 [4]/ViT-B/16 [3] image encoder as a frozen encoder backbone of U-Net [17]/SETR [26] and train the decoder.

3.2 Experimental Results

Results on Medical Image Classification Notably, the final results are based on ARSA with "merged bboxes", with a relevant explanation being provided in Ablation Study. As shown in Table 1, our framework achieves competitive performance across four datasets. Especially, on COVIDx with the novel disease "COVID-19", ASG shows significant improvements, highlighting its generalization ability. Fine-tuning with 1% of data, our ASG outperforms MGCA by 0.6% in AUR on NIH X-ray, demonstrating its ability to comprehend and distinguish a broader variety of diseases (see Supp. Fig. 2). Due to the prototype-level global clustering module in MGCA, it exhibits a slight advantage over our ASG on datasets with fewer categories, i.e., CheXpert and RSNA.

Results on Medical Semantic Segmentation Table 2 presents the semantic segmentation performance results on the SIIM and RSNA datasets. ASG demonstrates superior performance compared to all SOTA methods across every data fractions. Notably, ASG achieves a Dice score of 71.9% with only 1% data fine-tuning on the smaller-scale SIIM, surpasses the runner-up method by 3.6%, demonstrating the robust dense prediction capability of our framework.

Qualitative Analysis As shown in Fig. 3, to better understand the working mechanism of our ASG framework, we visualize the correspondence between images and disease words. ASG accurately highlights the relevant regions corresponding to a given disease, assisting the model in precise classification.

Table 2. Comparison with other SOTA methods on segmentation task.

Method	SIIM (Dice)			RSNA (Dice)		
	1%	10%	100%	1%	10%	100%
Random Init	9.00	28.6	54.3	6.90	10.6	18.5
ImageNet Init	10.2	35.5	63.5	34.8	39.9	64.0
CNN-based						
ConVIRT [25]	25.0	43.2	59.9	55.0	67.4	67.5
GLoRIA [5]	37.4	57.1	64.2	60.3	68.7	68.3
MedKLIP [22]	55.1	62.0	66.8	64.7	68.9	70.3
KAD [24]	58.4	68.2	69.9	67.9	68.5	70.3
MGCA [19]	49.7	59.3	64.2	63.0	68.3	69.8
Ours	60.7	66.7	73.6	68.4	69.9	72.6
ViT-based						
MRM [27]	68.3	69.5	72.2	69.5	69.2	70.6
MGCA [19]	60.1	65.4	69.6	69.3	70.0	72.3
Ours	**71.9**	**74.7**	**75.6**	**71.7**	**72.3**	**72.8**

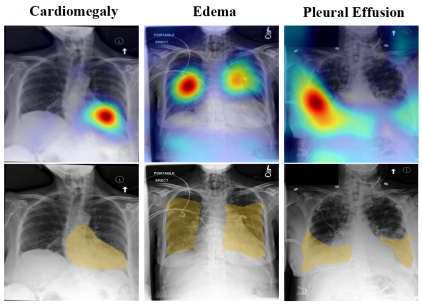


Fig. 3. Heat maps of the vision-language association learned by ASG, compared with GT annotations provided by radiologists.

Table 3. Ablation study of our framework.† and # denotes ARSA based on merge bbox and split sentence, respectively.

Learning Objective			NIH X-ray(AUC)			COVIDx(ACC)			RSNA(Dice)		
IRA	ARSA	IERL	1%	10%	100%	1%	10%	100%	1%	10%	100%
✓			78.2	81.7	82.6	75.3	85.8	91.0	65.1	67.7	68.3
✓	✓†		79.1	81.8	83.1	77.5	86.0	92.3	70.6	71.2	71.9
✓	✓#		78.9	81.5	83.4	76.3	86.3	92.0	69.0	69.4	69.7
✓		✓	78.8	81.7	83.4	79.3	86.5	92.8	67.4	68.6	69.7
✓	✓†	✓	**79.5**	**82.2**	**83.6**	**81.3**	**87.0**	**93.3**	**71.7**	**72.3**	**72.8**

Ablation Study We conducted the ablation study on two tasks with three datasets, the detailed results are shown in Table 3. The incorporation of the ARSA brings improvements in both classification and segmentation tasks, facilitating model to focus on local lesion representations across the entire image. ARSA based on merged bboxes outperforms that based on split sentences, likely because the former allows the model to learn the connections between different anatomical regions. The optimization of IERL is more pronounced in improving classification tasks performance, demonstrating a more reasonable approach to global representation modeling. Ultimately, the integration of these improvements yields the best overall performance.

4 Conclusion

We introduce a novel Anatomical Structure-Guided framework for medical vision-language pre-training. Firstly, we parse raw reports into triplets and then utilize each element. By aligning anatomical regions and sentences, we improve the model's localization ability and interpretability. The model's capabilities are further enhanced by improvements in both internal and external representation learning. In future, we will focus on further improving sentence parsing and anatomical region extraction accuracy, for more tasks, such as report generation.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Anna, Z., Carol, W., George, S., Julia, E., Mikhail, F., Mohannad, H., ParasLakhani, Phil, C., Shunxing, B.: Siim-acr pneumothorax segmentation (2019), https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
7. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
10. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
11. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
12. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
18. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
19. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems **35**, 33536–33549 (2022)
20. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific reports **10**(1), 19549 (2020)
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
22. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. medRxiv pp. 2023–01 (2023)
23. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316 (2021)
24. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
25. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
26. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
27. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. arXiv preprint arXiv:2301.13155 (2023)